PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Quality assessment of nutrition coverage in the media: A 6 week survey of five popular UK newspapers
AUTHORS	Kininmonth, Alice; Jamil, Nafeesa; Almatrouk, Nasser; Evans, Charlotte

VERSION 1 – REVIEW

REVIEWER	Catriona Bonfiglioli, PhD University of Technology Sydney Australia
REVIEW RETURNED	10-Nov-2016

GENERAL COMMENTS	This is an interesting and valuable paper which contributes to a scholarly discussion much in need of additional (and up-to-date) empirical evidence of this kind. I am not sufficiently expert to provide a review of the statistics, however, I note that Table 3 has two values in the "Mean" column each bearing a pair of asterisks. I assume that one of these two should have a single asterisk (perhaps the row for Obesity?). My key advice is that the description of the sampling strategy be edited to make clear how the researchers decided whether a newspaper article was about nutrition or not. Table 3 suggests that eligible articles are about many other key health issues apart from nutrition. I assume that the sampling strategy demanded that articles mention nutrition somewhere but I should not have to assume it should be clear in the method section. Rather than recognise that journalism -- even health reporting -- is a profession dedicated to reporting the news and not a public health information service, the authors cite research which criticises the media for how they classify stories as "newsworthy". Journalism certainly has professional goals to inform, educate, and entertain the public but, even before the internet, news has had to be structured to capture people's attention (sell newspapers). Health professionals and researchers would perhaps like news to be more of a handmaiden to health but news is a consumer-driven practice where newsworthiness is king. This is not to say health journalism cannot be improved only that analyses such as this need to acknowledge that journalists are bound by their own codes of ethics and professional values and success depends on providing audiences with news which attracts and holds their attention. The sampling strategy was driven by popularity (circulation) which left the researchers with a good sample of "tabloids" but only one "quality" newspaper. As tabloids tend to have shorter news stories and this study finds that shorter news stories are less likely to score well on their quality scale, this limitation should be acknowledged as shaping results. Where is The Times, The Guardian? The authors make some unsupported assertions: just because the writer receives a by-line does not mean they are a health journalist or even a journalist -- the sampling strategy does not explain how

the researchers made sure they were analysing news and not opinion columns (which could be written by a variety of non-journalists). Nor can we assume that because there is no by-line that the story was not written by a health reporter. The customs for assigning by-lines (or not) vary from publication to publication. The lack of a by-line could signify that the story was re-written from a press release or that it came from a wire news service or that it was written by a junior reporter or it was a minor story written by an established health reporter for whom it was not the significant story.

On page 18, line 52, the authors assert that "misrepresentation" occurred in their sample, however, I see no explicit evidence for this. If they can identify misrepresentation this should be supported with evidence, if not, please delete this.

While it's possible that no UK evidence has shown improvements in health reporting (MS page 19, line 12) in the last 20 years, Australian research led by Amanda Wilson (2009) used a systematic analytical instrument and found improvements in Australian health news. REF: Wilson, A., B. Bonevski, Jones, A & Henry, D. (2009). Media reporting of health interventions: signs of improvement, but major problems persist. PLoS ONE 4(3): e4831

While the Science Media Centre is certainly a frequently quoted source of science information it has links to commercial entities which should be acknowledged -- for more on this see: http://www.scidev.net/global/journalism/feature/uk-s-science-media-centre-lambasted-for-pushing-corporate-science.html

I look forward to seeing Dr Kininmonth and colleagues' revised manuscript.

Some line-by-line suggestions

page 2, line 4 (in the abstract) -- "the quality" - this phrase assumes that the instrument used is a sufficient measure of absolute "quality" -- suggest consider wording which acknowledges the study is measuring the quality of presentation of research and not a university quality or journalistic quality.

Page 2, line 4 (in the abstract) -- "top five" newspapers - please provide circulation figures and their source (at least in the main body of the paper).

Page 2, line 19/20 (abstract) -- "relevant" - this does not explain that stories were selected on the grounds of nutrition content.

Page 2, line 24 -- "poor quality" - consider re-wording such as "poor use of research to support assertions".

Page 2, lines 46/47 (abstract conclusions) -- "very poor quality" -- poor use of research to support assertions.

Page 3, line 4 -- Change "online newspapers" to "online news"

Page 4, lines 8/9 - suggest change "excess weight" to "weight gain" or "excess weight gain" -- obesity could be seen to be "excess weight"

Page 4, lines 46-48 -- The sentence citing ref.no.12 is too cryptic to be useful -- "newsworthiness" is a key determining factor for deciding whether to research and report a story/event/issue in what way did ref.12 criticise the "classification"?

Page 4, Line 53 -- "contradictory messages" -- "conflict" is a news value -- see Conley and Lamble's book The Daily Miracle.

Page 4, line 56-57 -- REF 15 discusses the autism-vaccine issue -- the news reporting of the Wakefield paper was significantly affected the key paper being published in the highly regarded medical

	journal The Lancet -- consider re-wording criticism of news media to acknowledge that they were reporting on a leading journal. Page 5, line 2-4-6 -- did the authors cited (REF 8) actually question the quality of a "majority" of the health claims? Page 5, line 15, delete full stop after newspapers and insert comma or semi-colon, otherwise the following sentence is not grammatical. Page 5, lines 26/27 and lines 30-33 - unqualified use of the word quality. Page 6, lines 6 and 7 -- no rationale is provided for choosing the top five circulation national papers and not the top six or seven. No circulation figures or sources are provided. The sampling strategy includes only one broadsheet (quality or elite) newspaper. This should be discussed as a limitation because tabloids tend to publish shorter stories than broadsheets and, as the authors themselves find, shorter stories are less able to include the elements which affect their "quality" rating -- where is The Guardian? The Times? Page 6, line 26 --Sunday papers probably have a different audience and could have been included. Page 6, line 29 -- change "research" to "researcher" Page 6, lines 29-38 -- this section should include a brief statement explaining how the researchers decided whether something was about nutrition AND human health -- what nutrition-related search terms made articles eligible? This study cannot be replicated without providing this detail. Page 6, line 38 -- "inclusion criteria" are not explicit. Please add detail. Page 6, line 53/54 -- while the authors cite The Eatwell Guide as providing categories they don't explain how or which categories. Page 7, line 11 -- authors refer to half page etc without specifying that they mean half of a tabloid page (assuming that all the "tabloid" newspapers were not Berliners https://www.theguardian.com/gpc/berliner-format or some other size) Page 7, line 40/41 -- "quality of reporting", "poor quality" -- acknowledge the instrument measures quality of use of evidence/academic papers. It's interesting that using the word "breakthrough" is considered a demerit despite some medical news being what the layperson -- and sometimes the scientific expert themselves -- regard as breakthrough. However, I acknowledge this is a word which should be used in moderation. Page 8, lines 1/2 -- not all stories published in newspapers are written by journalists. Did this study limit the sample to news articles? Newspapers also contain opinion pieces which can be written by almost anyone, journalist or not. Page 9, line 16-18 AND line 20 -- suggest change "publications" to articles or stories Page 9, line 40 -- delete apostrophe from "inches" Page 9, line 42 -- delete comma after the word Express. Page 9, line 47 -- suggest change "large-sized" to long. Articles are usually discussed as short or long, not big or small. Page 11, lines 7 and 18/19 -- "quality of reporting"; "high quality" -- see my comments above about "quality" Page 13, line 30 -- "size" --> "length" Page 13, line 37 -- "large" --> "long" Page 14, Table 3 -- check use of double asterisk -- perhaps the "obesity" line should have only one asterisk? Page 14, line 45-46 -- the authors note 70% of articles provide a

	second opinion -- this is a sign of high quality journalism. Fairness and balance are two values respected in journalism, featuring in codes of ethics and/or press council guidelines. Page 16, lines 6/7 and 12 -- "quality"; "poor quality". Page 16, line 16/17 -- here the authors assume that articles with by-lines are written by health journalists. Do they have any evidence of this? Looks like an unsupported assumption to me. Page 16, line 34-35 -- "varied in their interest to publish" -- suggest re-word to read: "varied in the number of nutrition-related articles they published" Page 17, line 14 -- "too blame" should read "to blame" Page 17, line 54 -- authors say they found news articles gave same weight to scholarly research whether it was an RCT or a cross-sectional study -- an important point but where is the evidence for this presented? Page 18, line 3 -- "misleading way" -- where are the quotes or other descriptions to support this assertion. Page 18, line 6-7 -- There are many resources for reporting on health, have the authors checked to see if any of them are nutrition related?: for example Moynihan, Soumerai and Bero's tipsheet can be found here: http://www.commonwealthfund.org/publications/press-releases/2001/mar/tipsheet-for-health-care-journalists-identifies-crucial-questions OR here: www.commonwealthfund.org and this website lists many resources for health journalism: http://discoverycentre.ru.ac.za/index.php?option=com_content&view=article&id=134&Itemid=197 Page 18, line 18-21 -- The authors do not present evidence to support the newspaper articles without by-lines (anonymous) would have been written by non-specialist journalists. It may be true but it looks like an assumption. Page 18, lines 27-28 -- "published" should read "written" -- journalists working for legacy media such as these newspapers are not usually seen as publishing (although it's possible they are publishing themselves online). Page 18, lines 30-34 -- no evidence is presented to support the assertion that reporters find it easier to report on obesity or that they are more familiar with it. Although it is true that obesity has been highlighted in the press since about 2002. Page 18, line 52 - no evidence of "misrepresentation" is presented in the data -- delete or support with evidence. Page 19, line 12 -- While it's possible that no UK evidence has shown improvements in health reporting in the past 20 years, Australian research led by Amanda Wilson (2009) used a systematic analytical instrument and found improvements in Australian health news. REF: Wilson, A., B. Bonevski, Jones, A & Henry, D. (2009). Media reporting of health interventions: signs of improvement, but major problems persist. PLoS ONE 4(3): e4831 Page 19, lines 23-25 -- journalists may ignore research suggesting audiences say they want to see only agreed upon findings because a/ scientists, experts and nutrition experts often disagree with each

	other and b/ by the time consensus is reached the information is old and thus not news -- newness is a key news value which makes information newsworthy. Page 20, line 21 -- While the Science Media Centre is certainly a frequently quoted source of science information its neutrality has been questioned and it has links to commercial entities which should be acknowledged -- for more on this see: http://www.scidev.net/global/journalism/feature/uk-s-science-media-centre-lambasted-for-pushing-corporate-science.html

REVIEWER	Dr Amy Nimegeer MRC/CSO Social and Public Health Sciences Unit, University of Glasgow, UK
REVIEW RETURNED	11-Nov-2016

GENERAL COMMENTS	This is a very interesting paper on scientific quality in nutrition coverage in the media and I enjoyed reading it. This paper, in my opinion, requires a few small adjustments prior to publication: - There is perhaps a lack of acknowledgement, particularly at the start of this paper, of social media as being part of the media landscape and a place where people encounter nutritional information (sometimes of dubious scientific quality). You do mention, when describing limitations, that people may be getting news from online newspapers or blogs but this would be helpful up front, as well as perhaps an acknowledgement of other social media such as facebook, twitter, etc. The UK still has a relatively high newspaper readership compared to other countries though so I think your focus on print journalism still has merit, but it is worth acknowledging the fuller picture of media influence. - In your Methods section, a sentence on inclusion or exclusion criteria for articles would be helpful in terms of replicability. - In your Methods section, you have described your strategy of sampling papers based on readership numbers, and I think this is entirely justifiable, however in your conclusions it might be worth making it clearer that the majority of newspapers that you sampled were tabloid/mid-market, so while this may represent a majority of media consumed, it does not necessarily represent the reporting habits of "any newspaper" (Discussion, Page 16, Line 10). For example, if your sample had included a wider range of papers including more broadsheets this may, or may not, have influenced your results. - In your Results section, Tables 2 and 3 the tables could be a little clearer. When there is a significant relationship denoted by asterisks, it would be nice to have a footnote explaining what the relationships are. - In the Results section (Page 13, Line 14) there should be no apostrophe in Thursdays or Tuesdays. - In the Discussion (Page 16, Line 30) you may want to provide evidence from the audience response literature that illustrates a link between poor quality reporting and readers becoming uninterested in a topic, if existing. - In the Discussion (Page 16, Line 45/46) you may want to provide

	evidence for (or maybe reword) the statement beginning “Articles are often published in newspapers”. While this assumption about editorial intent may be true, editors may also have other motivations for publishing particular articles. Additionally, you may want to touch on framing and agenda setting theories which consider the way in which the media ‘frames’ particular health topics, and how this can shape social norms/attitudes and therefore may influence audience understanding of and appetite for particular stories. I think this is important as it formulates part of your 'theory of change' for this article. - The reference to the Science Media Centre on Page 20, Line 21 feels slightly awkward as a conclusion, consider rewording?

REVIEWER	Ying Cao The State University of New York at Buffalo, USA
REVIEW RETURNED	06-Dec-2016

GENERAL COMMENTS	The study collected nutrition articles from 5 daily newspapers from UK for 6 weeks and investigated the article quality and patterns. After reading the manuscript, I have the following comments to share with the authors: 1. Study contribution This paper follows directly from Robinson et al. (2003), which used 8 UK newspapers for 4 weeks. This does not warrant enough contribution to the knowledge field. The authors mentioned that Robinson et al. covered shorter periods and hence, had smaller sample size, but no information was given on how the 141 articles in the current study is better in terms of sample size, not to mention with only 2 weeks longer, yet 3 newspaper less. Another concern is on how to frame the study purpose. The study used a standardized and validated tool to measure the quality score of each newspaper article, but then claimed to investigate what factors are driven these quality changes. First, if the quality scoring matrices are already selected, which means the authors agree that these matrices are comprehensive enough to objectively measure article quality, then by logic there should not exist other factors that would also driven the quality changes, otherwise, the quality measure system is failed in the first place. Alternatively, the study could claim to investigate the changing patterns of articles across newspaper and over time (i.e. weekly pattern, etc.). 2. Statistical analysis The analysis should go beyond single variable ANOVA, since the study testes many potential factors that contribute to the quality changing patters, such as article length, authorship, days of the week, and media, etc. These variables many also interact with one another, and hence, should be included in one multi-variable regression analysis to further control the variation and eliminating the confounding effects. 2.1 Article length and authorship It might be better to use the exact column width/word count to measure article length, which would yield more accurate measure and estimation. Yet, the findings with article length and anonymity of authorship do not add many scientific insights to the story. The fact that shorter and/or anonymous articles have lower quality may merely be a selection procedure from the editorial office due to their quality control. Are authors trying to suggest readers that do not read

	shorter article or anonymous ones? Or the editorial offices do not publish them? Neither of these is realistic. 2.2 Days of the week, week, topic and media The study found that days of the week have significant effects on article quality, with those published on Thursdays having higher quality. What is missed in the manuscript is some discussion about why this is the case and follow-up studies. For example, it could be that Thursday is approaching the latter part of the working week, when more life style articles could attract more attention. A further investigation should be performed on how article length and journal anonymity is correlated with days of the week. If Thursdays’ articles are longer and have definite authors, this will explain the earlier findings. Moreover, this correlation would also warrant a multivariate analysis rather than single ANOVA. The findings that weeks of the article have non-trend fluctuating effects could be due to the 6 weeks duration. While 4 weeks capture a month’s cycle, 8 weeks capture two and 12-13 weeks capture a season, 6 weeks might just be a period falls in between, and hence, hard to find any clear trend. The findings of quality variation across nutrition topics are interesting, with obesity as the most common concerns being of lowest quality. Yet, it is also suggested to double check with the article length, authorship and days of the week cycle. The reason is that with this being the most common topic, it is also the most easiest article topic to write a shorter one, by someone without professional background (so, anonymous) and during those busier days (such as Monday/Tue/Wed) when people are too much engaged to their working issues and have limited cognitive ability to fully digest other more sophisticated topics such as muscular skeletal, etc. The paper claimed to investigate the quality pattern across the 5 newspapers, but other than an overall test of equality in Table 2, there is no discussion on the issue. Pair-wise test might offer more information. 3. Figure presentation of the results To show the full distribution of the quality measures across media or across nutrition topics, a bar/line chart or stochastic dominance figure will be more straightforward and easier to compare.

VERSION 1 – AUTHOR RESPONSE

Comments

Reviewer 1

1

This is an interesting and valuable paper which contributes to a scholarly discussion much in need of additional evidence.

Response: We thank the reviewer for their positive feedback.

2

I am not sufficiently expert in statistics to provide a review of the statistics, however, I note that Table 3 has two values in the "Mean" column each bearing a pair of asterisks. I assume that one of these two should have a single asterisk (perhaps the row for Obesity?).

Response: We apologise for the confusion. The asterisks are removed from the tables for clarity as the results from the bonferoni test are reported clearly in the text.

3
My key advice is that the description of the sampling strategy be edited to make clear how the researchers decided whether a newspaper article was about nutrition or not. Table 3 suggests that eligible articles are about many other key health issues apart from nutrition. I assume that the sampling strategy demanded that articles mention nutrition somewhere but I should not have to assume -- it should be clear in the method section.

Response: We have strengthened the section in the methods and made it clear that included articles had to include an aspect of nutrition as an exposure and an aspect of health as a health outcome. We excluded anything in the opinion pages.
104-8

4
Rather than recognise that journalism -- even health reporting -- is a profession dedicated to reporting the news and not a public health information service, the authors cite research which criticises the media for how they classify stories as "newsworthy". Journalism certainly has professional goals to inform, educate, and entertain the public but, even before the internet, news has had to be structured to capture people's attention (sell newspapers). Health professionals and researchers would perhaps like news to be more of a handmaiden to health but news is a consumer-driven practice where newsworthiness is king. This is not to say health journalism cannot be improved only that analyses such as this need to acknowledge that journalists are bound by their own codes of ethics and professional values and success depends on providing audiences with news which attracts and holds their attention as well as maintaining high standards of research and reporting.

Response: The authors agree this is a very good point and the discussion has been strengthened to reflect this conflict of interest between news and health. A statement on how the role of newspapers is to report interesting news not to provide a public health service is inserted in the introduction and discussion. In addition the general tone of the discussion is checked to ensure that the discussion provides a more balanced argument.

5
The sampling strategy was driven by popularity (circulation) which left the researchers with a good sample of "tabloids" but only one "quality" newspaper. As tabloids tend to have shorter news stories and this study finds that shorter news

Response: This is a good point but actually there were more shorter articles in the Telegraph (the only broadsheet included here) than in the tabloids. The Times and the Guardian have
stories are less likely to score well on their quality scale, this limitation should be acknowledged as shaping results. Where is The Times? The Guardian?
smaller circulation numbers. Further analysis indicated that papers did vary in quality by article size and this is included in the results.

6
The authors make some unsupported assertions: just because the writer receives a by-line does not mean they are a health journalist or even a journalist -- the sampling strategy does not explain how the researchers made sure they were analysing news and not opinion columns (which could be written by a variety of non-journalists). Nor can we assume that because there is no by-line that the story was not written by a health reporter. The customs for assigning by-lines (or not) vary from publication to publication. The lack of a by-line could signify that the story was re-written from a press release or that it came from a wire news service or that it was written by a junior reporter or it was a minor story written by an established health reporter for whom it was not the significant story.

On page 18, line 52, the authors assert that "misrepresentation" occurred in their sample, however, I see no explicit evidence for this. If they can identify misrepresentation this should be supported with evidence, if not, please delete this.

Response: We have explained in the methods that opinion columns were not included in the search. Thank you for making these important points. It is true that the presence of bylines are down to editorial policy but the vast majority of newspapers in the UK now use them. However we have stated that newspapers may vary in the limitations.
We have made it clear in the discussion that an article could have been written by a health journalist even if no name is given and if no name given it could be more likely to be from a press release.
We have tightened up the wording on misrepresentation. The low quality score includes questions on misrepresentation and also there is cited evidence that misrepresentation exists and therefore we have still included this in the discussion.

7
While it's possible that no UK evidence has shown improvements in health reporting (MS page 19, line 12) in the last 20 years, Australian research led by Amanda Wilson (2009) used a systematic analytical instrument and found improvements in Australian health news. REF: Wilson, A., B. Bonevski, Jones, A & Henry, D. (2009). Media reporting of health interventions: signs of improvement, but major problems persist. PLoS ONE 4(3): e4831

Response: Thank you for this reference. We have incorporated it into the discussion.

8
While the Science Media Centre is certainly a frequently quoted source of science information its neutrality is a subject of debate and it has links to commercial entities which should be acknowledged -- for more on this see: http://www.scidev.net/global/journalism/feature/

Response: A sentence has been included to indicate that the Centre is not neutral.
328-330
uk-s-science-media-centre-lambasted-for-pushing-corporate-science.html

Reviewer: 2
9
This is a very interesting paper on scientific quality in nutrition coverage in the media and I enjoyed reading it.

Response: We thank the reviewer for the positive feedback.

10
- There is perhaps a lack of acknowledgement, particularly at the start of this paper, of social media as being part of the media landscape and a place where people encounter nutritional information (sometimes of dubious scientific quality). You do mention, when describing limitations, that people may be getting news from online newspapers or blogs but this would be helpful up front, as well as perhaps an acknowledgement of other social media such as facebook, twitter, etc. The UK still has a relatively high newspaper readership compared to other countries though so I think your focus on print journalism still has merit, but it is worth acknowledging the fuller picture of media influence.

Response: More detail has been added in the introduction on the role of social media in nutritional information.

11

- In your Methods section, a sentence on inclusion or exclusion criteria for articles would be helpful in terms of replicability.

Response: Please see point 3 above from reviewer 1

12
- In your Methods section, you have described your strategy of sampling papers based on readership numbers, and I think this is entirely justifiable, however in your conclusions it might be worth making it clearer that the majority of newspapers that you sampled were tabloid/mid-market, so while this may represent a majority of media consumed, it does not necessarily represent the reporting habits of "any newspaper" (Discussion, Page 16, Line 10). For example, if your sample had included a wider range of papers including more broadsheets this may, or may not, have influenced your results.

Response: This point has been added to the discussion (also see comment made in related point 5)

13
- In your Results section, Tables 2 and 3 the tables could be a little clearer. When there is a significant relationship denoted by asterisks, it would be nice to have a footnote explaining what the relationships are.

Response: The tables have been clarified by removing the asterisks. The results of the statistical analysis are clearly provided in the text.
Tables

14
- In the Results section (Page 13, Line 14) there should be no apostrophe in Thursdays or Tuesdays.

Response: These have been removed.

15
- In the Discussion (Page 16, Line 30) you may want to provide evidence from the audience response

Response: We have provided a reference for this statement.
literature that illustrates a link between poor quality reporting and readers becoming uninterested in a topic, if existing.

16
- In the Discussion (Page 16, Line 45/46) you may want to provide evidence for (or maybe reword) the statement beginning “Articles are often published in newspapers”. While this assumption about editorial intent may be true, editors may also have other motivations for publishing particular articles. Additionally, you may want to touch on framing and agenda setting theories which consider the way in which the media ‘frames’ particular health topics, and how this can shape social norms/attitudes and therefore may influence audience understanding of and appetite for particular stories. I think this is important as it formulates part of your 'theory of change' for this article.

Response: This paragraph has been reworded and the word ‘may’ inserted .
A sentence has been inserted on the influence of the media.

17
- The reference to the Science Media Centre on Page 20, Line 21 feels slightly awkward as a conclusion, consider rewording?

Response: The conclusion has been reworded and the SMC reference has been moved to earlier in the discussion.

Reviewer: 3
18
1. Study contribution
This paper follows directly from Robinson et al. (2003), which used 8 UK newspapers for 4 weeks. This does not warrant enough contribution to the knowledge field. The authors mentioned that Robinson et al. covered shorter periods and hence, had smaller sample size, but no information was given on how the 141 articles in the current study is better in terms of sample size, not to mention with only 2 weeks longer, yet 3 newspaper less.

Response: We thank the reviewer for these important points.
The sample size for both studies is similar and this is now clarified in the text.
It is important that this manuscript adds to what is known already. We agree that a regression analysis is needed to determine the most important factors adjusted for other factors. See also point .

19
Another concern is on how to frame the study purpose. The study used a standardized and validated tool to measure the quality score of each newspaper article, but then claimed to investigate what factors are driven these quality changes. First, if the quality scoring matrices are already selected, which means the authors agree that these matrices are comprehensive enough to objectively measure article quality, then by logic there should not exist other factors that would also driven the quality changes, otherwise, the quality measure system is failed in the first place. Alternatively, the study could claim to investigate

Response: The authors agree it is important to clarify the purpose. We are looking at factors that explain differences in quality of articles between papers. These factors are listed in the aims. The multiple regression strengthens the findings.

the changing patterns of articles across newspaper and over time (i.e. weekly pattern, etc.).
20
2. Statistical analysis
The analysis should go beyond single variable ANOVA, since the study testes many potential factors that contribute to the quality changing patters, such as article length, authorship, days of the week, and media, etc. These variables many also interact with one another, and hence, should be included in one multi-variable regression analysis to further control the variation and eliminating the confounding effects.

Response: The authors agree regression analysis is warranted here to provide more informative results. Multiple regression has been carried out and results reported in a new table (table 3). This changed the final results slightly but not markedly. The whole manuscript has been updated to reflect these changes.
throughout

21
2.1 Article length and authorship
It might be better to use the exact column width/word count to measure article length, which would yield more accurate measure and estimation.
Yet, the findings with article length and anonymity of authorship do not add many scientific insights to the story. The fact that shorter and/or anonymous articles have lower quality may merely be a selection procedure from the editorial office due to their quality control. Are authors trying to suggest

readers that do not read shorter article or anonymous ones? Or the editorial offices do not publish them? Neither of these is realistic.
The authors believe that there is no universally agreed method for measuring length of article and therefore have used this method but state in the limitations that there may be measurement error issues which could be addressed in future.

Response: We thank the reviewer for raising important points. We believe these are useful points to raise to start a discussion on how to improve the quality of nutrition in the media. We agree that this needs to be tackled in association with newspapers for common ground to be found.

2.2 Days of the week, week, topic and media The study found that days of the week have significant effects on article quality, with those published on Thursdays having higher quality. What is missed in the manuscript is some discussion about why this is the case and follow-up studies. For example, it could be that Thursday is approaching the latter part of the working week, when more life style articles could attract more attention. A further investigation should be performed on how article length and journal anonymity is correlated with days of the week. If Thursdays' articles are longer and have definite authors, this will explain the earlier findings. Moreover, this correlation would also warrant a multivariate analysis rather than single ANOVA.
The findings that weeks of the article have non-trend fluctuating effects could be due to the 6 weeks duration. While 4 weeks capture a month's cycle, 8 weeks capture two and 12-13 weeks

Response: We agree the different factors need exploring.
We have stated in the limitations that we are not able to validly look at fluctuations by week.
More rigorous analysis has been presented on the exploration of the differences between papers and how the differences are explained by different factors. Information has been added to the aims, methods, results and discussion.

This also includes some results on obesity and other factors. The authors believe that the power is not sufficient to explore these correlations between variables formally but data has
capture a season, 6 weeks might just be a period falls in between, and hence, hard to find any clear trend.
The findings of quality variation across nutrition topics are interesting, with obesity as the most common concerns being of lowest quality. Yet, it is also suggested to double check with the article length, authorship and days of the week cycle. The reason is that with this being the most common topic, it is also the most easiest article topic to write a shorter one, by someone without professional background (so, anonymous) and during those busier days (such as Monday/Tue/Wed) when people are too much engaged to their working issues and have limited cognitive ability to fully digest other more sophisticated topics such as muscular skeletal, etc.
The paper claimed to investigate the quality pattern across the 5 newspapers, but other than an overall test of equality in Table 2, there is no discussion on the issue. Pair-wise test might offer more information.
been provided on obesity and predictor variables including day of week and anonymity. There was little difference by article length.

3. Figure presentation of the results
To show the full distribution of the quality measures across media or across nutrition topics, a bar/line chart or stochastic dominance figure will be more straightforward and easier to compare.

Response: The authors believe that the tables are the most informative way of providing the information and have provided a new table of the multiple regression results

VERSION 2 – REVIEW

REVIEWER	Catriona Bonfiglioli University of Technology Sydney, Australia
REVIEW RETURNED	11-Feb-2017

GENERAL COMMENTS	Thanks for the opportunity to see the revised manuscript. I look forward to seeing the published paper. I note that I am not in a position to review the additional statistical analysis.

REVIEWER	Dr Amy Nimegeer MRC/CSO Social and Public Health Sciences Unit, University of Glasgow, UK
REVIEW RETURNED	23-Feb-2017

GENERAL COMMENTS	Thank you for your amendments, the tables in particular seem clearer to me know (admittedly I am a non-statistician!) My remaining feedback relates predominantly to the Discussion section. Although you begin to acknowledge it, I still think that the limitations section does not fully capture the fact that the papers selected (top 6 by readership) may not convey the breadth of reporting on nutrition in the UK during the period of the study. I think it is important to emphasise this because you are drawing conclusions about the quality of reporting in newspapers in the UK and your sample excludes several broadsheets that may change the distribution of your results (i.e. are perhaps more likely to contain longer, more descriptive articles). Just a bit more emphasis needed. In line 260 of the Discussion, you claim "this is the first study that explains differences in article quality between newspapers". This is inaccurate and needs to be more specific to accurately reflect your study contribution (i.e. is it the first article in the UK, looking at scientific quality of nutrition reporting?) Also, I don't think that the study explains the differences so much as describes them so you may want to clarify that. To this end, you do make some causal inferences that perhaps exceed the evidence provided, such as that differences in quality are due to editorial policies of the papers (there is no evidence provided regarding specific editorial policies and their relation to the articles). I would possibly reword to suggest that possibly this may be a contributory factor? Similarly with lines 291-294 in the Discussion section - I think this might be too forcefully stated, what makes you think this might be the case? Lines 265-267 in the Discussion, you claim that "Journalists may perceive that it is easier to write a news article on obesity than heart disease as they feel more familiar with the subject" - the way this is written at the moment, I would think requires evidence of the link between familiarity with a subject and number of articles written/desire to write on a topic. Might it be similarly plausible that they think obesity has a high prevalence therefore might be of personal interest to more readers than heart disease? Lines 359-360 “there has been little improvement to the quality of reporting nearly 30 years later”: I think you are making quite a definitive claim here about changes over time based on a very small and non-comparable sample.

REVIEWER	Regina L. Nuzzo Gallaudet University, USA
REVIEW RETURNED	22-May-2017

GENERAL COMMENTS

This is an interesting paper in an area of wide interest and importance. My comments below will focus on major statistical aspects of the manuscript.

4. The statistical methods described in lines 140 – 153 are not sufficient to allow other researchers to fully understand and reproduce the analyses reported on here. Since the bulk of the authors' arguments and conclusion rest on the results obtained from statistical analyses, it is important that the authors expand the Statistical Analysis section to provide more details about exactly what methodology was used and how it was used in each of Tables 1 – 4. Alternatively, the authors could provide more details about each analysis as it is discussed in the Results: Quality Assessment section.
7. I have questions and concerns about the statistical methodology used for Tables 1 – 4.

General comments about the statistical analysis description in Lines 144 – 146: A one-way ANOVA is an analysis for a quantitative outcome and a single categorical predictor; i.e., it compares the means of some measure for two or more groups.

Therefore, it can be used "to compare quality of reporting across the five newspapers," where newspaper is the single categorical predictor. But a one-way ANOVA cannot be used "to determine which of the six factors listed above individually influenced article quality." For the multiple regression model, it's not clear how determining "differences in quality score between newspaper title when adjusted for all other predictors" will help "thereby determine which were the key predictors of quality." Are the authors analysing nested models? Comparing goodness of fit? It's not clear.

Table 1: The authors report mean and standard deviation, but a glance at the descriptive statistics suggests that the quality scores might be skewed. If so, median and IQR would be appropriate descriptive measures of center and spread to report, and the use of a 95% CI might not be appropriate. If there is room, boxplots of quality scores for each newspaper would be instructive. Also, the information about column inches and article size is fairly redundant – the article size appears to just be a categorization of the quantitative measure of column inches. Is a categorization even necessary? What extra information does it add? A better use of space in that table would be to explain the categorization of quality score that the authors mentioned in lines 172 – 174. Why were these cutoffs for "poor," "satisfactory," and "high" chosen? What is the frequency breakdown of poor, satisfactory, and high quality articles for each of the five newspapers?

Lines 176 – 178: P-values of the differences between quality of reporting in the newspapers was provided, but it's not clear how these p-values were obtained. Was this from a one-way ANOVA with quality score as the outcome and newspaper title as the predictor factor? If so, where is the overall p-value for the factor? If post-hoc Bonferroni tests were conducted, this means that there should be 10 pairwise comparisons, but only 4 were reported (and it was not mentioned whether these p-values were unadjusted or adjusted for multiple comparisons). It goes against best practices to cherry-pick comparisons to report, or to set up a reference category after viewing the data.

Lines 180 – 184: Again, it’s not clear how this p-value were obtained. If these are the results from a one-way ANOVA with quality score as the outcome and week as the predictor factor, then there should be a p-value for the overall effect of week. If post-hoc pairwise comparisons using Bonferroni correction were used as declared in the statistical analysis section, then there should be 15 pairwise comparisons with p-values. It appears that Week 1 was compared to results in Weeks 2 – 6 combined. If so, and unless there was an a priori reason (before viewing the data) to choose Week 1 as a baseline reference point, using this grouping to do a statistical comparison is a “questionable research practice” often known as p-hacking, and is statistically unethical.

Lines 184 – 187: Again, it’s not clear whether these are the results of an ANOVA. They appear to be a single pairwise comparison (Thursday vs Tuesday) that was decided upon after viewing the data. Following what the authors said was their analysis plans, there should be a one-way ANOVA to determine the significance of the effect of day (with a reported p-value of the effect), and then 15 pairwise Bonferroni comparisons. Cherry-picking is not appropriate.

Lines 194 – 196: Exact p-values to three decimal places should be reported, not to two decimal places (unless it is BMJ style to do otherwise).

Lines 205 – 209: It’s not clear if this first p-value of 0.03 is from an ANOVA with quality score as the outcome and categorical article size (small, medium, large) as the predictor. The authors then report p-values from two pairwise comparisons (medium vs small, medium vs large) but omit the third pairwise comparison of small vs large. It’s also not clear whether these p-values are post hoc Bonferroni comparisons.

Table 2/Lines 217 – 218 and 219 - 220: In the previous sentence authors reported what appears to be a p-value for the effect of health category on quality score, from a one-way ANOVA. According to their analysis plan, they performed post hoc pairwise comparisons with Bonferroni corrections among the 8 health categories, which would result in 24 pairwise comparisons. The authors reported the results from only a single pairwise comparison (obesity vs CVD), and a contrast between obesity vs everything else. Reporting the results from only a single pairwise comparison is not full transparency and is not best practices for multiple comparisons reporting.
Table 3: The authors appear to have built a multiple regression model using dummy variables for the categorical predictors of newspaper, week, day, food, health, and byline. It’s not clear whether article size was treated as a categorical variable or a quantitative variable. The most appropriate model here would be to do a linear mixed effects model, with both random and fixed effects. The authors should consult a statistician or statistical textbook for more information about linear mixed models, and the difference between fixed effects (factors with levels deliberately chosen to be of interest, such as byline [present or absent]) and random effects

	(factors with levels that are not deliberately chosen to be of interest and are a random subset of a much larger set of levels, such as week [there is nothing special about weeks 1 through 6, and they are a subset of all weeks in the year]). Even if the results in Table 3 were obtained through a linear mixed model, the authors' method for concluding significance is in error. In lines 226 – 229, the authors say that the effect of newspaper is not significant in the full model. But to conclude this the authors need to test nested models – that is, comparing a full model with newspaper dummy variables included to a reduced model with newspaper dummy variables omitted. What the authors have done here is essentially to do pairwise comparisons between The Sun and each of the four other newspapers. This is the same problem as noted above – there needs to be an investigation of overall effect for the newspaper factor. The authors should again consult a statistician or statistical textbook about testing nested models with polytomous factors using dummy variables. Table 4: The authors appear to have done 21 different chi-square analyses on related data, which again leads to a multiple comparisons problem. The p-values appear not to have been adjusted for multiple comparisons.It should be noted that even if there were no true differences within the 21 tests, we would expect 5% of the p-values to be less than 0.05 simply by random chance, which would yield one or two p-values to be incorrectly statistically significant; the results indeed showed two statistically significant p-values. The authors might benefit from examining the literature and methods for False Discovery Rate correction.

REVIEWER	Dr Golnaz Shahtahmassebi Nottingham Trent University, UK
REVIEW RETURNED	30-May-2017

GENERAL COMMENTS	The manuscript has investigated very important topic. The discussion was extensive and interesting. However, the design of the study and statistical methods used in the manuscript require further revision. Please see my detailed comments below: 1. Selection of sources is biased: a. Considering advances in the technology and having access to internet almost everywhere in the UK it is vital to explore online articles. This important fact was discarded in the manuscript, in this way a large proportion of reader population has been ignored. Considering the fact that Robinson et. al. (2013) had conducted a similar study, therefore it was important to include online articles rather than printed ones only. b. It is very important to discuss the type audience for each newspaper, also number of reads for health related articles in these newspapers. Is the level of exposure to health related articles for these newspapers same as other articles? c. Also, it is worth to highlight each newspaper's policy with regards to fact checks and the level of accuracy they may consider in general. 2. It is not clear whether scoring articles was blinded or not. 3. Express the name of the Quality Assessment Tool. 4. Statistical analysis: a. The outcome measure was categorised on page 7 intro three categories: poor, satisfactory and high qualities. However, these categories were not analysed in any part of the article. I recommend

	to use this as the outcome measure and conduct independence test between the name of the newspaper and quality category. Then, depending on your contingency table, you might be able to use Pearson chi-square test. Some result was provided on page 10, which is not satisfactory. The count for each category-newspaper was not provided. b. The range of scores are between -9 to 10, without clear idea on the distribution of the data. There were 44 articles with score of less than zero and 97 articles with score of greater than 0. This can suggest the data was highly skewed to the left. In addition, the quality score for each newspaper is highly fluctuating. That is the quality score of articles within each newspaper is not homogenous which suggests using one way ANOVA is not appropriate for this study. c. This is also applicable for the regression analysis. Multiple linear regression (MLR) can only be used if the quality score data follows a normal distribution. My recommendation is to use non-parametric techniques, such as Kruskal-Wallis instead of one way ANOVA and nonparametric regression instead of standard MLR. You may also explore multinomial logistic regression if using quality scores as a categorical variable.

VERSION 2 – AUTHOR RESPONSE

Comment: The statistical methods described in lines 140 – 153 are not sufficient to allow other researchers to fully understand and reproduce the analyses reported on here in tables 1 to 4

Response: The statistical methods and results have been updated to clarify the methods used and results obtained.

Results
General comments about the statistical analysis description in Lines 144 – 146: A one-way ANOVA is an analysis for a quantitative outcome and a single categorical predictor; i.e., it compares the means of some measure for two or more groups. Therefore, it can be used “to compare quality of reporting across the five newspapers,” where newspaper is the single categorical predictor. But a one-way ANOVA cannot be used “to determine which of the six factors listed above individually influenced article quality.” For the multiple regression model, it’s not clear how determining “differences in quality score between newspaper title when adjusted for all other predictors” will help “thereby determine which were the key predictors of quality.” Are the authors analysing nested models? Comparing goodness of fit? It’s not clear. All anovas have been removed. Regression models were used for individual predictors and all predictors. A test of the overall model for each regression analysis was reported. The methods have been re-written to provide greater clarity. 144-163
Table 1: The authors report mean and standard deviation, but a glance at the descriptive statistics suggests that the quality scores might be skewed. If so, median and IQR would be appropriate descriptive measures of center and spread to report, and the use of a 95% CI might not be appropriate. We agree that the distribution of the quality scores is not perfect but it is broadly symmetrical and not skewed. The histogram of the residuals of the full model is normally distributed. There is no evidence of non-constant variance. (see plots at end of this document).

Response: A sentence has been added to the results.

Comment: If there is room, boxplots of quality scores for each newspaper would be instructive.

Response: All the information is given in the tables so we feel that boxplots are not needed here. n/a

Comment: Also, the information about column inches and article size is fairly redundant – the article size appears to just be a categorization of the quantitative measure of column inches. Is a categorization even necessary? What extra information does it add?

Response: Column inches have been removed. They are not normally distributed so are not informative.

Comment: A better use of space in that table would be to explain the categorization of quality score that the authors mentioned in lines 172 – 174. Why were these cutoffs for “poor,” “satisfactory,” and “high” chosen? What is the frequency breakdown of poor, satisfactory, and high quality articles for each of the five newspapers?

Response: The cut offs were set by Robinson et al who validated the tool. The frequency breakdown of poor and satisfactory quality articles have been added to table 1

Lines 176 – 178: P-values of the differences between quality of reporting in the newspapers was provided, but it’s not clear how these p-values were obtained. Was this from a one-way ANOVA with quality score as the outcome and newspaper title as the predictor factor? If so, where is the overall p-value for the factor? If post-hoc Bonferroni tests were conducted, this means that there should be 10 pairwise comparisons, but only 4 were reported (and it was not mentioned whether these p-values were unadjusted or adjusted for multiple comparisons). It goes against best practices to cherry-pick comparisons to report, or to set up a reference category after viewing the data.   p values for the overall models are reported throughout and the methods clarified so this is clear.

Response: No Bonferroni tests were included and these are all removed. Comparison with the reference was tested with the reference category being the category with the lowest score.

Lines 180 – 184: Again, it’s not clear how this p-value were obtained. If these are the results from a one-way ANOVA with quality score as the outcome and week as the predictor factor, then there should be a p-value for the overall effect of week. If post-hoc pairwise comparisons using Bonferroni correction were used as declared in the statistical analysis section, then there should be 15 pairwise comparisons with p-values.

Response: See section 7 above.144-163

Comment: It appears that Week 1 was compared to results in Weeks 2 – 6 combined. If so, and unless there was an a priori reason (before viewing the data) to choose Week 1 as a baseline reference point, using this grouping to do a statistical comparison is a “questionable research practice” often known as p-hacking, and is statistically unethical. The category with the lowest quality score was used throughout as the reference category. This decision was made a priori and is a consistent approach. It also provides the easiest interpretation for readers. We prefer this to comparing each category with every other category as this would introduce more multiple testing. Although our focus is generally on the estimates and confidence intervals rather than the p values.

Response: We have made it clear in the methods that all categories have been compared to the reference group. (note week was not included as a fixed effect.)   144-163

Lines 184 – 187: Again, it’s not clear whether these are the results of an ANOVA. They appear to be a single pairwise comparison (Thursday vs Tuesday) that was decided upon after viewing the data.

Following what the authors said was their analysis plans, there should be a one-way ANOVA to determine the significance of the effect of day (with a reported p-value of the effect), and then 15 pairwise Bonferroni comparisons. Cherry-picking is not appropriate.

Response: The anova has been removed. All categories of day have been compared with Tuesday and all results are reported whether significant or not.

Lines 194 – 196: Exact p-values to three decimal places should be reported, not to two decimal places (unless it is BMJ style to do otherwise).

Response: We have been advised from an experienced statistician that the important thing is to quote exact p values rather than categorised p-values (which is what we have done) so you can see how close or far away from 0.05 they are. However, P-values themselves are very unreliable, and would have very large prediction intervals around them if we were to show them. Therefore although exact p-values are needed, the imprecision in their estimation means that it is better to round them quite a lot. Often thinking in terms of significant figures makes more sense than decimal places for p-values. The authors therefore believe it is appropriate to report p values to 2dp in this case as it is easier to read. Other bmj open articles have also provided p values to 2dp so it is not against journal policy.

Lines 205 – 209: It's not clear if this first p-value of 0.03 is from an ANOVA with quality score as the outcome and categorical article size (small, medium, large) as the predictor. The authors then report p-values from two pairwise comparisons (medium vs small, medium vs large) but omit the third pairwise comparison of small vs large. It's also not clear whether these p-values are post hoc Bonferroni comparisons.

Response: See previous comments. The test for overall model is reported. Pairwise comparisons are deleted. Categories are collapsed from 3 to 2.

Table 2/Lines 217 – 218 and 219 - 220: In the previous sentence authors reported what appears to be a p-value for the effect of health category on quality score, from a one-way ANOVA. According to their analysis plan, they performed post hoc pairwise comparisons with Bonferroni corrections among the 8 health categories, which would result in 24 pairwise comparisons. The authors reported the results from only a single pairwise comparison (obesity vs CVD), and a contrast between obesity vs everything else. Reporting the results from only a single pairwise comparison is not full transparency and is not best practices for multiple comparisons reporting.

Response: See previous comments.

Table 3: The authors appear to have built a multiple regression model using dummy variables for the categorical predictors of newspaper, week, day, food, health, and byline. It's not clear whether article size was treated as a categorical variable or a quantitative variable.

Response: Article size is now a binary categorical variable (small and larger). Changed throughout the methods and results

Comment: The most appropriate model here would be to do a linear mixed effects model, with both random and fixed effects. The authors should consult a statistician or statistical textbook for more information about linear mixed models, and the difference between fixed effects (factors with levels deliberately chosen to be of interest, such as byline [present or absent]) and random effects (factors with levels that are not deliberately chosen to be of interest and are a random subset of a much larger set of levels, such as week [there is nothing special about weeks 1 through 6, and they are a subset of all weeks in the year]).

Response: We agree that the weeks could be different if another sample of weeks were sampled so have re-analysed with articles clustered within weeks. On advice from a statistician there are not enough weeks (clusters) to estimate the between-cluster variation. Newspapers (and everything else) are fixed, and week is random but we have used the sandwich estimator to take account of the hierarchical structure instead.

The methods have also been updated. The results in table 3 are updated.

Even if the results in Table 3 were obtained through a linear mixed model, the authors' method for concluding significance is in error. In lines 226 – 229, the authors say that the effect of newspaper is not significant in the full model. But to conclude this the authors need to test nested models – that is, comparing a full model with newspaper dummy variables included to a reduced model with newspaper dummy variables omitted. What the authors have done here is essentially to do pairwise comparisons between The Sun and each of the four other newspapers. This is the same problem as noted above – there needs to be an investigation of overall effect for the newspaper factor. The authors should again consult a statistician or statistical textbook about testing nested models with polytomous factors using dummy variables. A likelihood ratio test was used to test whether the model with newspaper in significantly improves the fit of the model compared with not having newspaper in. it was not a significant predictor of quality

Table 4: The authors appear to have done 21 different chi-square analyses on related data, which again leads to a multiple comparisons problem. The p-values appear not to have been adjusted for multiple comparisons. It should be noted that even if there were no true differences within the 21 tests, we would expect 5% of the p-values to be less than 0.05 simply by random chance, which would yield one or two p-values to be incorrectly statistically significant; the results indeed showed two statistically significant p-values. The authors might benefit from examining the literature and methods for False Discovery Rate correction.

Response: ''We agree with the reviewer that there is too much testing although these were purely exploratory not hypothesis confirming. So that a reader does not focus on any one significant result we have removed the testing and commented qualitatively on where main differences lie.

Reviewer: 2

Comment: The limitations section does not fully capture the fact that the papers selected (top 6 by readership) may not convey the breadth of reporting on nutrition in the UK during the period of the study. Important to emphasise this because you are drawing conclusions about the quality of reporting in newspapers in the UK and your sample excludes several broadsheets that may change the distribution of your results (i.e. are perhaps more likely to contain longer, more descriptive articles). Just a bit more emphasis needed.

Response: The authors believe they have made this point in the limitations. A whole paragraph discusses how this only captures a limited picture of all news sources. However, we have tried to emphasise this point further by strengthening the wording in some of the sentences.

Comment: In line 260 of the Discussion, you claim "this is the first study that explains differences in article quality between newspapers". This is inaccurate and needs to be more specific to accurately reflect your study contribution (i.e. is it the first article in the UK, looking at scientific quality of nutrition reporting?)

Response: This is the first study to explore in detail the predictors of article quality for nutrition related articles and this is now clarified in the first sentence.

Comment: Also, I don’t think that the study explains the differences so much as describes them so you may want to clarify that.

Response: We have changed from explain to describe the differences. Explain is a statistical term to indicate that percent variation has been accounted for by the variables so it has not been changed everywhere.
Comment: To this end, you do make some causal inferences that perhaps exceed the evidence provided, such as that differences in quality are due to editorial policies of the papers (there is no evidence provided regarding specific editorial policies and their relation to the articles). I would possibly reword to suggest that possibly this may be a contributory factor? Similarly with lines 291-294 in the Discussion section - I think this might be too forcefully stated, what makes you think this might be the case?

Response: We have changed the sentence to reflect that this is our opinion: These differences in article quality could possibly be related to editorial policy and other factors that were not considered here.

We have removed this comment as it is speculation: We did not collect relevant information to determine why quality of articles varied by day and perhaps the reasons for this need to be explored further.

Lines 265-267 in the Discussion, you claim that "Journalists may perceive that it is easier to write a news article on obesity than heart disease as they feel more familiar with the subject" - the way this is written at the moment, I would think requires evidence of the link between familiarity with a subject and number of articles written/desire to write on a topic. Might it be similarly plausible that they think obesity has a high prevalence therefore might be of personal interest to more readers than heart disease?

Response: We agree that we have no evidence for this and it has been removed. n/a
Lines 359-360 “there has been little improvement to the quality of reporting nearly 30 years later”: I think you are making quite a definitive claim here about changes over time based on a very small and non-comparable sample.
This has been deleted and replaced with: many of these issues still persist.

Reviewer 1
Comment: Selection of sources is biased. It is vital to include online sources.

Response: We agree that online sources are extremely important. The tool was designed to assess quality of articles in printed newspapers but could be used with online sources. We collected the data in 2014 and did not include them. This is a major limitation of the data and we have expressed this clearly in the limitations. We stated that further research in this area must include online sources.

Comment: Important to discuss the type of audience for each newspapers, also the number of reads for health related articles in these newspapers. Is the level of exposure to health related articles for these newspapers the same as other articles.

Response: The definition of tabloid newspaper includes the type of audience that is most likely to be attracted to the paper. We have included a reference to the definitions of tabloid and broadsheet.

We have already included information that readers are interested in health.

Comment: Highlight each newspaper's policy with regards to fact checks and the level of accuracy they may consider in general.

Response: We agree that this would be interesting but this information is difficult to find for printed newspapers although we are aware that the Daily Mail has been accused of publishing too much fake news.
Comment: Not clear whether scoring articles was blinded or not   The scoring was done in duplicate without knowing the other person's score. The word 'independently' scored was inserted 112
Give the name of the Quality Assessment Tool   No other name is given for the quality assessment tool Outcome measure of quality score was categorised into poor, satisfactory and good but not in the article. The count for each category for each newspaper was not provided

Response: This information is now added to table 1.

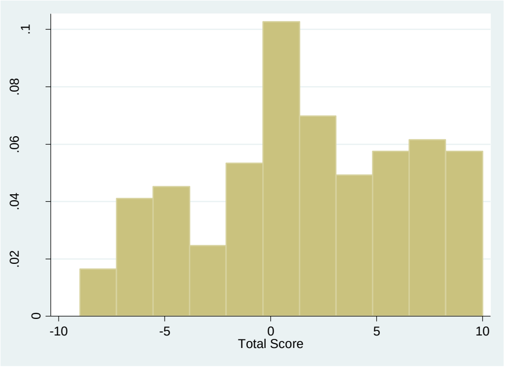
Comment: After seeking advice from a statistician we have made the decision to use quality score as a continuous variable (see point) Distribution of scores suggests they are skewed and vary within newspaper

Response: We agree that the quality scores do not have a perfect distribution. However they are broadly symmetrical and not skewed (see plot at the end of this table). Anovas have been removed and replaced with regression techniques. The residuals are normally distributed (see plot) and importantly the variance is constant.

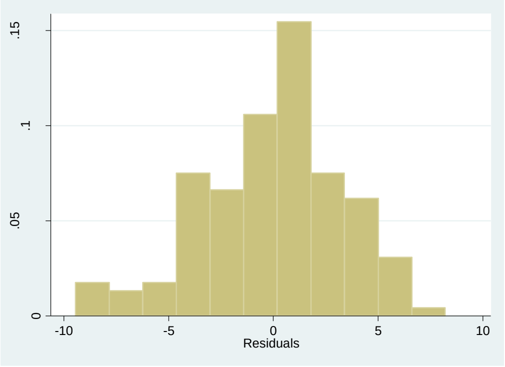
Comment: Regression analysis. Quality score needs to follow a normal distribution. Non parametric methods may be appropriate.

Response: The advice of an experienced statistician is that the variable does not need to be normally distributed but the unexplained variation does. The residuals indicate that this is the case and non – parametric methods are not needed.

Histogram of quality score



Histogram of residuals of multiple regression model



VERSION 3 – REVIEW

REVIEWER	Regina Nuzzo Gallaudet University, U.S.A.
REVIEW RETURNED	04-Aug-2017

GENERAL COMMENTS	The authors are to be commended for a much improved statistical analysis. I have a few remaining questions and concerns: -- In line 34, the authors state that the articles were generally of poor quality. However the mean quality score (and all values in its 95% CI) is considered “satisfactory,” not “poor” according to Robinson’s measure. It may be on the low end of this category, but if the authors use Robinson’s classification then they need to do so faithfully. Indeed in Table 1, it can be seen that the majority (69% overall) of articles is of “satisfactory” quality. -- Why do the results presented in Table 3 contradict those presented in the "Quality assessment" section? -- What is the justification for running a regression for each predictor independently when there is likely to be confounding relationships among the predictors? Separate regressions may give the bivariate relationships between each predictor and the response, but the statistical significance and R^2 is likely to be misleading if considered independently. -- The difference in interpretation between results from a separate simple regression models and results from a full regression model should be made very clear. One accounts for confounders; the other does not. -- In lines 185 - 194, it should be made clear that the authors looked at only a subset of comparisons among newspapers rather than all the pairwise comparisons. Stating that "there were differences between . . some of the individual newspapers" and following up with some comparisons implies that the authors are reporting on all comparisons, which isn't true. That is, the analysis was set up to look at whether the Sun differed from any of the other newspapers, and from this analysis a significant difference was found only between the Sun and the Daily Telegraph and between the Sun and the Daily Express. No conclusions can be made about whether, for example, Daily Mail articles differed from Daily Telegraph. Unless there was a scientific reason a priori (BEFORE looking at the

results) to set the Sun as the reference group, reporting this subset of comparisons is a form of cherry-picking, and readers should be aware of this.

-- It’s not clear why it’s important to report the results from the comparison of week 1 to each of the other weeks. Week is a random effect here, and these 6 weeks were chosen at random from the 52 weeks of the year. Week 1 was chosen as a baseline after looking at the data, and there is no reason to think that there is an interesting difference between Week 1 and Week 2, and so on. It is sufficient to note the significant random effect of weeks.

-- For day of week analysis, Tuesday was chosen as a baseline reference after looking at the data. Again, these are not all the pairwise comparisons. If the authors think that the confidence interval for the quality difference between Tuesday and Thursday is important to report, why would the quality difference between Thursday and Monday not be important to report? There is nothing a priori interesting about Tuesday, as far as I can tell, so I don’t understand why the authors are not reporting all the pairwise comparisons (some of which may be statistically significant).

-- For article size analysis, in lines 225 - 232, three categories are used, and a statistically significant difference was found in quality among the categories. Yet the authors then combine the medium and large categories into one “larger” category without a priori justification. The authors say there was no consistent trend by size, but a U-shaped trend is still a trend, especially if the overall effect is statistically significant. The authors also say that medium and large article “were similar,” but similar in what way? In terms of their average quality? Again, decisions to categorize data should be made a priori or with scientific reasons, not after looking at the data.

-- In line 257 the authors say that a sandwich estimator was used. A sandwich estimator of what? Why is this important? As this is a non-standard statistical element, a reference should be provided.

-- In lines 263 - 264 the authors say that “food type was also not an important predictor.” If I understand the authors’ analysis correctly, for determining the significance of paper type, they did a nested model approach, comparing a full model with a model with all predictors except paper type. This is indeed the appropriate approach. But the authors did not appear to do the same thing for food type. In order to make a conclusion about the overall effect of food type when all other factors are held constant, the authors should have compared a full model to a nested model with all predictors minus food type. Simply looking at the significance of the dummy variables compared to a single baseline will not reveal the significance of the effect of food type. The same thing applies for the effect of day of week and the effect of health category.

REVIEWER	Dr Golnaz Shahtahmassebi Nottingham Trent University, UK
REVIEW RETURNED	25-Aug-2017

GENERAL COMMENTS	The manuscript has improved considerably. My only concern related to modelling quality scores remains. They still show strong skewness to the left. Some more goodness of fit is required. A histogram is not enough to show the normality of errors. Still, working with score categories looks more appropriate.

VERSION 3 – AUTHOR RESPONSE

Reviewer 4

1 In line 34, the authors state that the articles were generally of poor quality. However the mean quality score (and all values in its 95% CI) is considered “satisfactory,” not “poor” according to Robinson’s measure. It may be on the low end of this category, but if the authors use Robinson’s classification then they need to do so faithfully. Indeed in Table 1, it can be seen that the majority (69% overall) of articles is of “satisfactory” quality.

Response: The abstract has been updated and this statement has been replaced with the percent of poor articles (31%).

2 Why do the results presented in Table 3 contradict those presented in the "Quality assessment" section? We have added a sentence to say that different papers performed badly on particular questions but no newspaper fared badly on all questions.

Response: We do not believe that the results contradict those presented earlier.

3 What is the justification for running a regression for each predictor independently when there is likely to be confounding relationships among the predictors? Separate regressions may give the bivariate relationships between each predictor and the response, but the statistical significance and R^2 is likely to be misleading if considered independently.

Response: The authors agree that separate regressions for each predictor are not needed and they have been removed. The statistical methods have been re-written. lines 143-63

4 The difference in interpretation between results from a separate simple regression models and results from a full regression model should be made very clear. One accounts for confounders; the other does not.

Response: See comment above. The simple regressions are no longer included except for newspaper type as we were interested in differences between papers when no predictors were included.

5 In lines 185 - 194, it should be made clear that the authors looked at only a subset of comparisons among newspapers rather than all the pairwise comparisons. Stating that "there were differences between some of the individual newspapers" and following up with some comparisons implies that the authors are reporting on all comparisons, which isn't true. That is, the analysis was set up to look at whether the Sun differed from any of the other newspapers, and from this analysis a significant difference was found only between the Sun and the Daily Telegraph and between the Sun and the Daily Express. No conclusions can be made about whether, for example, Daily Mail articles differed from Daily Telegraph. Unless there was a scientific reason a priori (BEFORE looking at the results) to set the Sun as the reference group, reporting this subset of comparisons is a form of cherry-picking, and readers should be aware of this.

Response: The authors agree that we can’t cherry pick but having failed to make a priori decisions we need to take a standard approach. Making the category with the highest frequency the reference category is a standard approach. The reference category is therefore the Daily Mail newspaper which had the most nutrition articles, Tuesday for day of the week, obesity for health outcome, energy for food type and yes for whether there was a by-line.

See results section.
6 It’s not clear why it’s important to report the results from the comparison of week 1 to each of the other weeks. Week is a random effect here, and these 6 weeks were chosen at random from the 52 weeks of the year. Week 1 was chosen as a baseline after looking at the data, and there is no reason to think that there is an interesting difference between Week 1 and Week 2, and so on. It is sufficient to note the significant random effect of weeks.
Any comparison of weeks has been removed as we were not specifically interested in weeks. We have said that there is significant variation between weeks.

7 For day of week analysis, Tuesday was chosen as a baseline reference after looking at the data. Again, these are not all the pairwise comparisons. If the authors think that the confidence interval for the quality difference between Tuesday and Thursday is important to report, why would the quality difference between Thursday and Monday not be important to report? There is nothing a priori interesting about Tuesday, as far as I can tell, so I don’t understand why the authors are not reporting all the pairwise comparisons (some of which may be statistically significant).

Response: We have explained that we have used the most common category as the reference category (see point 5). We do not want to report all pairwise comparisons due to multiple testing pointed out by reviewer 5 in the second round of corrections.

8 For article size analysis, in lines 225 - 232, three categories are used, and a statistically significant difference was found in quality among the categories. Yet the authors then combine the medium and large categories into one “larger” category without a priori justification. The authors say there was no consistent trend by size, but a U-shaped trend is still a trend, especially if the overall effect is statistically significant. The authors also say that medium and large article “were similar,” but similar in what way? In terms of their average quality? Again, decisions to categorize data should be made a priori or with scientific reasons, not after looking at the data.

Response: We agree we should not have done this. Article size has been maintained with the original 3 categories.

9 In line 257 the authors say that a sandwich estimator was used. A sandwich estimator of what? Why is this important? As this is a non-standard statistical element, a reference should be provided.

Response: This is a standard method used in statistics. A reference has been added.

10 In lines 263 - 264 the authors say that “food type was also not an important predictor.” If I understand the authors’ analysis correctly, for determining the significance of paper type, they did a nested model approach, comparing a full model with a model with all predictors except paper type. This is indeed the appropriate approach. But the authors did not appear to do the same thing for food type. In order to make a conclusion about the overall effect of food type when all other factors are held constant, the authors should have compared a full model to a nested model with all predictors minus food type. Simply looking at the significance of the dummy variables compared to a single baseline will not reveal the significance of the effect of food type. The same thing applies for the effect of day of week and the effect of health category.

Response: We have repeated the nested model approach for all the predictors and have reported the results (replacing the simple regressions with these results instead). see Results section

Reviewer 5

11 My only concern related to modelling quality scores remains. They still show strong skewness to the left. Some more goodness of fit is required. A histogram is not enough to show the normality of errors. Still, working with score categories looks more appropriate.

Response: We have changed the analysis to logistic regression using a binary score (not satisfactory/satisfactory quality score). Interpretation is generally similar although article size was no longer statistically significant. Changes have been made to the abstract, methods, results and the discussion accordingly.

VERSION 4 – REVIEW

REVIEWER	Regina Nuzzo Gallaudet University, USA
REVIEW RETURNED	30-Sep-2017

GENERAL COMMENTS	This is a greatly improved manuscript and statistical analysis from the previous submission. Just a few comments: 1) The authors mention on line 291 that the lowest quality score articles were published on Tuesday. But Table 3 shows the odds ratio of Saturday compared to Tuesday as being less than 1, which means that Saturday has the lowest-quality articles, not Tuesday. 2) I appreciate that the authors are concerned about multiple post-hoc comparisons. But the problem with multiple comparisons is when all the comparisons are not fully reported, or when their p-values are interpreted in isolation without recognition of the family-wise error rate. Transparent reporting and/or p-value adjustments are a reasonable solution, instead of stopping short and not providing more detailed information about main effects that are statistically significant. There are many statistical methods that will allow researchers to conduct post-hoc multiple comparisons and get p-values that will reveal individual differences while also controlling for false positives. False Discovery Rate adjustments are one good option, but there are a host of pairwise comparison methods available in standard statistical packages that will work as well, including Bonferroni and Scheffe methods. If the authors prefer not to take advantage of this, that's their choice, but it does limit the usefulness of information they provide. For example, in the text of the results in lines 241 - 248, the authors omitted discussion of the significant main effect of Food. I presume this is because their limited post-hoc pairwise comparisons (Energy vs each of the other 6 categories) did not reveal any significant differences. Yet we know from the Likelihood Ratio Test of the nested models that some pairwise differences in Food categories are contributing significantly to the model. It seems like this would be the appropriate time to delve further to see which of those categories is driving the significant impact, for better or for worse (is it alcohol? high fat and processed foods?), instead of just ignoring the effect entirely. 3) As possible explanation for differences in quality between days, the authors mention that perhaps different journals prefer to disseminate press releases on some days more than others. I'm not sure if the authors are aware of the embargo policy that most journals have, which restrict publication before a certain day of the week, which is usually consistent for each journal. It seems that embargo policies rather than dissemination of press releases will dictate when newspapers publish their articles.

REVIEWER	Dr Golnaz Shahtahmassebi Nottingham Trent University
REVIEW RETURNED	16-Oct-2017

GENERAL COMMENTS	Changes made are appropriate.

VERSION 4 – AUTHOR RESPONSE

1. The authors mention on line 291 that the lowest quality score articles were published on Tuesday. But Table 3 shows the odds ratio of Saturday compared to Tuesday as being less than 1, which means that Saturday has the lowest-quality articles, not Tuesday.

Response: lines 244-9

Tuesday is the lowest score when not adjusted for other factors but when adjusted for other factors Saturday has the lowest score. A sentence has been included in the results to suggest that other factors may be more common on Saturdays driving the typical score down (although there were no obvious differences for Saturday articles so it could be due to as yet unknown factors). In the discussion the mention of Tuesday has been removed to avoid confusion.

2. I appreciate that the authors are concerned about multiple post-hoc comparisons. But the problem with multiple comparisons is when all the comparisons are not fully reported, or when their p-values are interpreted in isolation without recognition of the family-wise error rate. Transparent reporting and/or p-value adjustments are a reasonable solution, instead of stopping short and not providing more detailed information about main effects that are statistically significant. There are many statistical methods that will allow researchers to conduct post-hoc multiple comparisons and get p-values that will reveal individual differences while also controlling for false positives. False Discovery Rate adjustments are one good option, but there are a host of pairwise comparison methods available in standard statistical packages that will work as well, including Bonferroni and Scheffe methods. If the authors prefer not to take advantage of this, that's their choice, but it does limit the usefulness of information they provide. For example, in the text of the results in lines 241 - 248, the authors omitted discussion of the significant main effect of Food. I presume this is because their limited post-hoc pairwise comparisons (Energy vs each of the other 6 categories) did not reveal any significant differences. Yet we know from the Likelihood Ratio Test of the nested models that some pairwise differences in Food categories are contributing significantly to the model. It seems like this would be the appropriate time to delve further to see which of those categories is driving the significant impact, for better or for worse instead of just ignoring the effect.

Response: lines in Methods 149-151, 160, results 193, 241-254

Pairwise comparisons have now been carried out with a Bonferroni correction. Information is included in the methods and results. One result was statistically significant (days of the week: Thursday vs Saturday). All other comparisons were not significant. A sentence on food categories was included to clarify this result.

3. As possible explanation for differences in quality between days, the authors mention that perhaps different journals prefer to disseminate press releases on some days more than others. I'm not sure if the authors are aware of the embargo policy that most journals have, which restrict publication before a certain day of the week, which is usually consistent for each journal. It seems that embargo policies rather than dissemination of press releases will dictate when newspapers publish their articles.

Response: line 332

We agree this is important. The statement on press releases has been replaced with information on embargo policies.

VERSION 5 – REVIEW

REVIEWER	Regina Nuzzo Gallaudet University, U.S.A.
REVIEW RETURNED	08-Nov-2017

GENERAL COMMENTS	Thank you for your responsiveness to feedback and your diligence in producing a manuscript with very solid methodology and conclusions. Nicely done.